Quantifying the Rewiring of Non-Linear Property Couplings Across Molecular Scaffolds via Mantel Tests

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Molecular properties are frequently modeled as independent targets, obscuring the intrinsic, multivariate joint structures that govern the physical state space of small molecules. These underlying relationships are often masked by dominant stoichiometric dependencies. To isolate genuine chemical organization from trivial size effects, we analyze 133,802 molecules from the QM9 dataset by regressing out exact atom counts using non-linear XGBoost models. We partition the molecules into acyclic, saturated cyclic, and aromatic cyclic scaffolds, and compute robust distance correlation matrices on size-balanced, bootstrapped residuals. We then employ Mantel tests and Procrustes analysis to statistically quantify how the multivariate property spaces deform across these discrete architectural boundaries. Our results demonstrate that non-linear models successfully expose hidden chemical signals, revealing that while acyclic and saturated cyclic manifolds share similar property topologies, the introduction of aromaticity fundamentally rewires the joint physical structure. Specifically, aromaticity elevates electronic properties to central hubs within the property network and decouples classical geometric-vibrational axes. Additionally, we find that geometrically inconsistent molecules exhibit abnormally rigid joint structures, suggesting they represent specific strained configurations rather than mere computational artifacts.

1 Introduction

The physical state space of small molecules is governed by complex, multivariate relationships among thermodynamic, electronic, and geometric properties. In computational chemistry and machine learning, these properties are frequently modeled as independent prediction targets. While this isolated treatment simplifies model optimization, it obscures the intrinsic joint structures that define molecular behavior, limiting our understanding of how fundamental physical traits co-vary across different chemical regimes.

A major challenge in mapping this multivariate space is the overwhelming influence of stoichiometric determinism. Trivial size effects, driven by exact atom counts, dominate the variance of many molecular properties and mask more subtle, genuine chemical organization. Traditional attempts to control for molecular size often rely on linear models, which fail to capture the non-linear scaling of geometric and spatial properties. Consequently, residual size-leakage persists, confounding the analysis of how structural features—such as the formation of rings or the onset of electronic delocalization—affect the underlying property manifold.

To rigorously isolate genuine chemical signals from stoichiometric dependencies, this study introduces a statistical framework based on non-linear baseline removal and multivariate distance metrics. We analyze 133,802 molecules from the QM9 dataset, regressing out exact atom counts using non-linear XGBoost models to fully strip the size-leakage that linear models fail to capture. By partitioning the dataset into discrete architectural classes—specifically acyclic, saturated cyclic, and aromatic cyclic scaffolds—we establish a controlled environment to observe how the physical state space is rewired by fundamental structural shifts.

Within size-balanced and bootstrapped samples of these structural classes, we compute robust distance correlation matrices of the non-linear residuals. This approach captures core non-linear joint structures, such as geometric-vibrational axes and rotational-spatial couplings, entirely independent of molecular size. To statistically quantify how these multivariate property spaces deform across chemical boundaries, we employ Mantel tests and Procrustes analysis. Additionally, we isolate latent geometrically inconsistent molecules to

determine whether their divergent behavior stems from underlying computational artifacts or represents a distinct, highly strained chemical regime.

Our results demonstrate that non-linear stoichiometric neutralization successfully exposes hidden topological features of the molecular property space. We reveal that while acyclic and saturated cyclic manifolds share remarkably similar property topologies, the introduction of aromaticity fundamentally rewires the joint physical structure. In aromatic systems, electronic properties are elevated to central hubs within the property network, and classical geometric-vibrational axes are decoupled. Furthermore, we show that geometrically inconsistent molecules exhibit abnormally rigid joint structures, suggesting they represent specific strained configurations rather than mere noise. By quantifying these scaffold-specific topological deformations, this work provides a statistically validated measure of how architectural shifts reorganize the joint physical state space of small molecules.

2 Methods

2.1 Data preprocessing and structural classification

To establish a robust foundation for analyzing the multivariate property space, we utilized the QM9 dataset, deduplicating entries based on SMILES strings to yield 133,802 unique molecules. We isolated 12 non-redundant physical properties to prevent trivial rank-deficiencies in downstream matrix operations: the rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), highest occupied molecular orbital energy ($homo$), lowest unoccupied molecular orbital energy ($lumo$), electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), internal energy at 0 K (u_0), heat capacity at 298.15 K (c_v), and atomization energy at 0 K (u_{0_atom}). The HOMO-LUMO gap (gap) was additionally retained for a controlled sensitivity analysis, bringing the working set to 13 properties. Because the rotational constant A exhibits an extreme heavy-tailed distribution (reaching up to 619,867 GHz) and is exactly zero for linear molecules, we applied a $\log(A+1)$ transformation to robustly scale the values without distorting their relative ordering.

Molecular structures were parsed to extract exact atom counts (C, N, O, F, and H), total heavy-atom counts, molecular weights, and ring counts. Linear molecules were explicitly flagged using a binary indicator where $A = 0$ and $B = C$. Molecules were then partitioned into three broad architectural categories based on their Murcko scaffolds: Acyclic (0 rings), Saturated Cyclic (containing rings, but no aromatic atoms or bonds), and Aromatic Cyclic (containing at least one aromatic ring). To rigorously isolate the effects of electronic delocalization from geometric constraints, we identified matched pairs of saturated and aromatic scaffolds that share the exact same heavy-atom skeleton, heavy-atom count, and ring count. Only scaffolds with a minimum sample size of $N \geq 500$ were retained to ensure stable statistical estimation.

2.2 Stoichiometric baseline removal

As outlined in the introduction, stoichiometric determinism heavily masks intrinsic chemical organization. To isolate genuine structural signals, we regressed out the exact atom counts (C, N, O, F, H) and the linear molecule flag from each of the 13 properties. For each property, we trained two separate models: a baseline Ordinary Least Squares (OLS) linear regression and a non-linear XGBoost regressor. By comparing the variance explained between the two approaches, we quantified the extent of non-linear size-leakage that traditional linear models fail to capture. The residuals from both models were retained to comparatively evaluate their impact on the underlying property manifold.

2.3 Outlier regime analysis

The QM9 dataset contains a known subset of geometrically inconsistent molecules that can act as high-leverage outliers. To determine whether these molecules represent mere density functional theory (DFT) convergence artifacts or a distinct, highly strained chemical regime, we performed an Isolation Forest analysis on the OLS property residuals, flagging the top 5% as multivariate outliers. We then computed and compared global distance correlation matrices for the inlier and outlier populations. A drastic divergence in the coupling structure would indicate that these geometrically inconsistent molecules occupy a fundamentally rigidified region of the physical state space.

2.4 Size-balancing and robust standardization

To ensure that observed deformations in the property manifold were not artifacts of differing stoichiometric distributions across scaffold classes, we implemented a rigorous size-balancing protocol. For each pairwise scaffold comparison, we drew 100 bootstrapped samples such that the heavy-atom and hydrogen count distributions of the two classes matched exactly. The bootstrap sample size was set to $N = \min(500, \text{total_scaffold_count})$. Crucially, to prevent extreme residual values from dominating the distance metrics, we applied a robust standardization—scaling by the interquartile range—to the residuals. This standardization was performed strictly *after* residual calculation and *within* each specific scaffold class and bootstrap sample.

2.5 Distance correlation and manifold topology

To capture the core non-linear joint structures of the property space, we computed pairwise distance correlation ($dCor$) matrices for each bootstrapped, standardized sample. A robust mean $dCor$ matrix was then derived for each structural class by taking the median across the 100 bootstrapped matrices. To verify that the inclusion of the redundant *gap* variable did not fundamentally alter the manifold topology, this entire procedure was performed twice: once on the 12 non-redundant properties and once on the full 13-property set.

2.6 Statistical quantification of manifold deformation

We employed Mantel tests with 10,000 permutations to statistically quantify how the multivariate property spaces deform across discrete chemical boundaries. We compared the robust mean $dCor$ matrix of the Acyclic baseline to those of the Cyclic scaffolds. Controlled comparisons were also executed between matched saturated and aromatic counterparts to isolate the specific rewiring driven by aromaticity. Furthermore, we compared the Mantel r-values derived from OLS versus XGBoost residuals to rigorously assess the necessity of non-linear stoichiometric neutralization.

To visualize and quantify these topological shifts, the robust mean $dCor$ matrices were converted into distance matrices using $1 - dCor$ and subjected to Multi-Dimensional Scaling (MDS). We performed Procrustes analysis (with optimal scaling enabled) to align the MDS embedding of each Cyclic scaffold's property space to the Acyclic baseline, extracting per-property Procrustes residuals to measure individual property displacement. Finally, we calculated the degree centrality for each property within the $dCor$ networks and computed the Pearson correlation between the property-specific Procrustes residuals and their absolute change in centrality, thereby identifying which physical properties emerge as central hubs during structural rewiring.

3 Results

3.1 The stoichiometric baseline: linear versus non-linear determinism

As outlined in our methodological framework, the most dominant factor governing the QM9 property space is molecular size, driven by exact atom counts. A baseline Ordinary Least Squares (OLS) linear regression on these counts recovers almost all variance in thermodynamic properties, notably internal energy at 0 K (u_0, $R^2 = 1.0000$), zero-point vibrational energy ($zpve$, $R^2 = 0.9971$), and atomization energy (u_{0_atom}, $R^2 = 0.9889$).

However, comparing these linear baselines to non-linear XGBoost models reveals the hidden chemical signals that linear stoichiometric neutralization fails to capture (). For the geometric rotational constants B and C, the linear model performs remarkably poorly ($R^2 < 0.07$), whereas the XGBoost model achieves an $R^2 \approx 0.945$. This 88% increase in variance explained indicates that rotational constants are fundamentally non-linear functions of stoichiometry, likely mediated by the spatial arrangement and branching of the atoms. In contrast, electronic properties ($homo$, $lumo$, gap, μ) remain difficult to explain even with non-linear models, yielding XGBoost R^2 values ranging from 0.198 to 0.483. This suggests that electronic features are driven by fine-grained topological details and specific bonding environments rather than mere elemental composition.

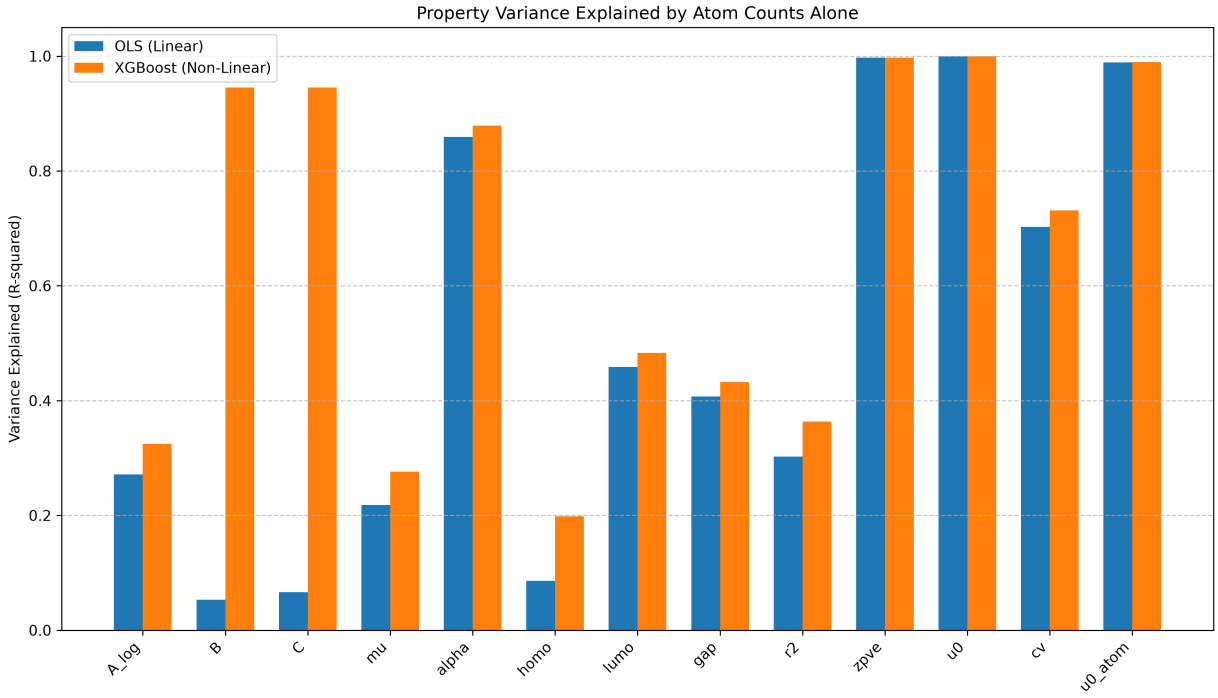

Figure 1: Variance explained (R^2) by atom counts for 13 molecular properties using linear (OLS) and non-linear (XGBoost) models. The comparison reveals that thermodynamic properties such as u_0 and *zpve* are almost entirely determined by linear stoichiometry. Conversely, geometric rotational constants (B and C) exhibit poor linear fits but achieve high accuracy with XGBoost, indicating an 88% increase in variance explained and demonstrating they are fundamentally non-linear functions of composition. Meanwhile, electronic properties (*homo*, *lumo*, *gap*, μ) remain poorly explained by both models, suggesting these features are driven by fine-grained topological details rather than mere stoichiometric composition.

3.2 Global joint structure and basis sensitivity

With size-leakage robustly neutralized via the XGBoost residuals, we mapped the core joint structure of the physical state space using distance correlation ($dCor$). Across the global dataset, properties exhibit a median off-diagonal $dCor$ of approximately 0.28, confirming the presence of a structured, multivariate manifold rather than a set of independent targets.

Our sensitivity analysis confirms that the inclusion of the redundant *gap* variable (the exact difference between *lumo* and *homo*) introduces a significant, but isolated, rank-deficiency without destabilizing the broader network. Mantel tests comparing the 12-property basis (excluding *gap*) to the 13-property basis demonstrate high topological stability ($r = 0.8742$ for OLS vs. XGBoost residuals in the acyclic set). Within this size-independent manifold, the joint structure is dominated by several key couplings. Most notably, acyclic molecules exhibit a strong geometric-vibrational axis between res_zpve and res_cv ($dCor = 0.7679$), as well as a tight rotational-spatial coupling between res_r^2 and res_C ($dCor = 0.7157$).

3.3 Structural rewiring across architectural families

Molecular architecture—specifically the introduction of rings and electronic delocalization—fundamentally reorganizes how these physical properties relate to one another. Mantel tests confirm that the acyclic and saturated cyclic manifolds share relatively similar topologies ($r = 0.7991$). However, the transition to aromatic cyclic structures causes the greatest deformation in the property space, yielding a Mantel correlation of only $r = 0.6931$ between the saturated and aromatic manifolds.

3.3.1 Procrustes deformation and centrality shifts

Multi-Dimensional Scaling (MDS) combined with Procrustes analysis allows for the visualization and quantification of this manifold deformation. provides an overview of the acyclic baseline embedding alongside the structural rewiring induced by aromaticity.

To track the specific property trajectories during this transition, we aligned the acyclic and aromatic manifolds (). The dipole moment (μ) exhibits the highest Procrustes residual (0.5664) during the transition from acyclic to aromatic architectures, indicating that its coupling to the rest of the property space is the most profoundly altered by the presence of delocalized rings.

Concurrently, as detailed in the centrality analysis (), the electronic orbital energies *homo* and *gap* show the largest increases in network centrality (+0.8749 and +0.5648, respectively). This shift indicates that in aromatic systems, the electronic state is elevated to a central hub that mediates other physical properties—a structural role it does not hold in acyclic or saturated cyclic molecules.

3.3.2 Coupling rewiring and difference heatmaps

Difference heatmaps of the $dCor$ matrices reveal the specific pairwise rewiring events that drive these global deformations when comparing aromatic molecules to their matched saturated counterparts ().

Three primary modes of rewiring emerge: 1. **Electronic Decoupling:** The coupling between *lumo* and *gap* drops significantly ($\Delta dCor = -0.4660$), while the *homo-lumo* coupling increases ($\Delta dCor = +0.4273$). 2. **Geometric Sensitivity:** The relationship between the rotational constants (B, C) and isotropic polarizability (α) weakens considerably in aromatic systems (e.g., B-α $\Delta dCor = -0.3754$). 3. **Vibrational Damping:** The tight link between electronic spatial extent (r^2) and heat capacity (c_v), which is prominent in acyclic and saturated systems, is heavily disrupted in aromatic rings ($\Delta dCor = -0.3691$).

3.4 Regime analysis: the geometric inconsistency effect

Finally, we investigated the 3,054 molecules previously flagged in the QM9 dataset for geometric inconsistency. Isolation Forest analysis on the multivariate residuals successfully identifies these molecules as high-leverage outliers. Rather than presenting as random computational noise, these outliers exhibit a vastly more rigid joint structure than the inlier population, with a mean off-diagonal $dCor$ of 0.4580 compared to 0.2849 for the inliers. Most strikingly, the *zpve*-c_v coupling in these outliers reaches an extreme $dCor = 0.9171$.

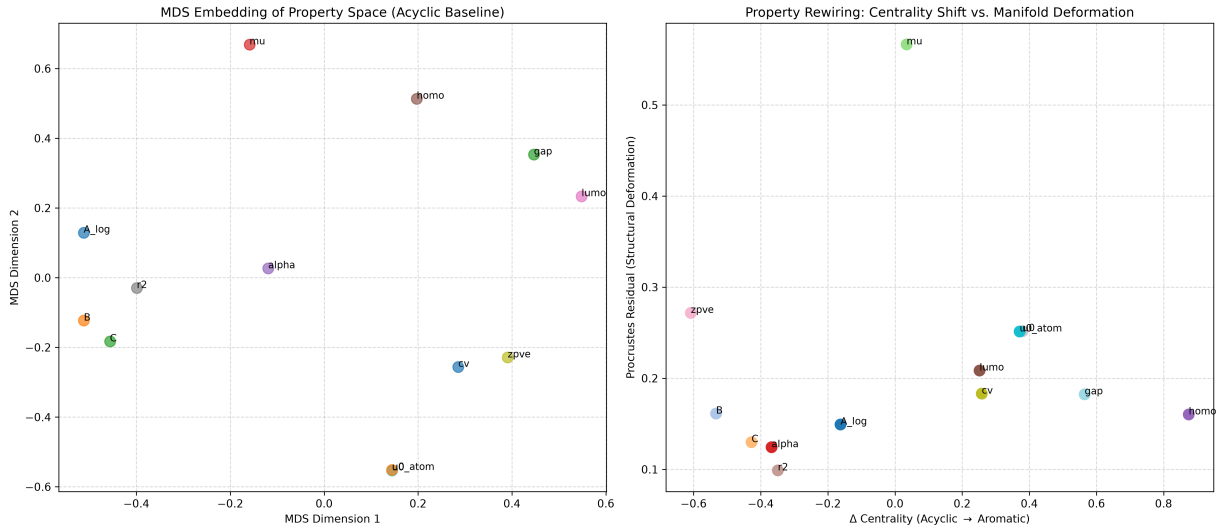

Figure 2: The left panel displays a Multi-Dimensional Scaling (MDS) embedding of the molecular property space for the acyclic baseline, mapping the initial joint structure and relationships between physical properties. The right panel illustrates the structural rewiring of this space during the transition from acyclic to aromatic architectures by plotting the Procrustes residual (structural deformation) against the shift in network centrality. As aromaticity is introduced, the dipole moment (μ) undergoes the highest structural deformation, indicating that its coupling to the rest of the property space is profoundly altered. Concurrently, electronic properties such as *homo* and *gap* exhibit the largest increases in centrality. This reveals that in aromatic systems, electronic states are elevated to central hubs that mediate other physical properties, fundamentally reorganizing the molecular state space.

This hyper-rigid topology suggests that these geometric inconsistencies do not merely stem from DFT convergence artifacts, but rather represent a distinct chemical regime of highly strained configurations where thermodynamic and geometric properties are abnormally forced into alignment.

4 Conclusions

The physical state space of small molecules is frequently approached as a collection of independent prediction targets, an assumption that obscures the intrinsic, multivariate joint structures governing molecular behavior. By employing non-linear XGBoost models to regress out exact atom counts, this study successfully isolated genuine chemical organization from the dominant masking effect of stoichiometric determinism. We demonstrated that while linear models adequately capture the size dependence of thermodynamic properties, they fundamentally fail to account for the non-linear scaling of geometric and spatial features, necessitating a more rigorous approach to baseline removal.

With size-leakage robustly neutralized, our distance correlation and manifold analyses revealed that molecular architecture fundamentally reorganizes the joint property space. While acyclic and saturated cyclic manifolds exhibit remarkably similar topologies, the introduction of aromaticity acts as a primary rewiring agent that drives a profound structural deformation. In aromatic systems, electronic properties such as *homo* and *gap* are elevated to central hubs within the property network. Concurrently, classical relationships—such as the geometric-vibrational link between electronic spatial extent (r^2) and heat capacity (c_v), as well as the sensitivity between rotational constants and polarizability—are significantly decoupled.

Furthermore, our multivariate residual framework provided new insights into the nature of geometrically inconsistent molecules within the QM9 dataset. Rather than presenting as random density functional theory convergence noise, these molecules occupy a distinct, hyper-rigid region of the physical state space. Their abnormally strong property couplings, particularly the extreme alignment between zero-point vibrational energy (*zpve*) and heat capacity (c_v), suggest they represent specific, highly strained structural configura-

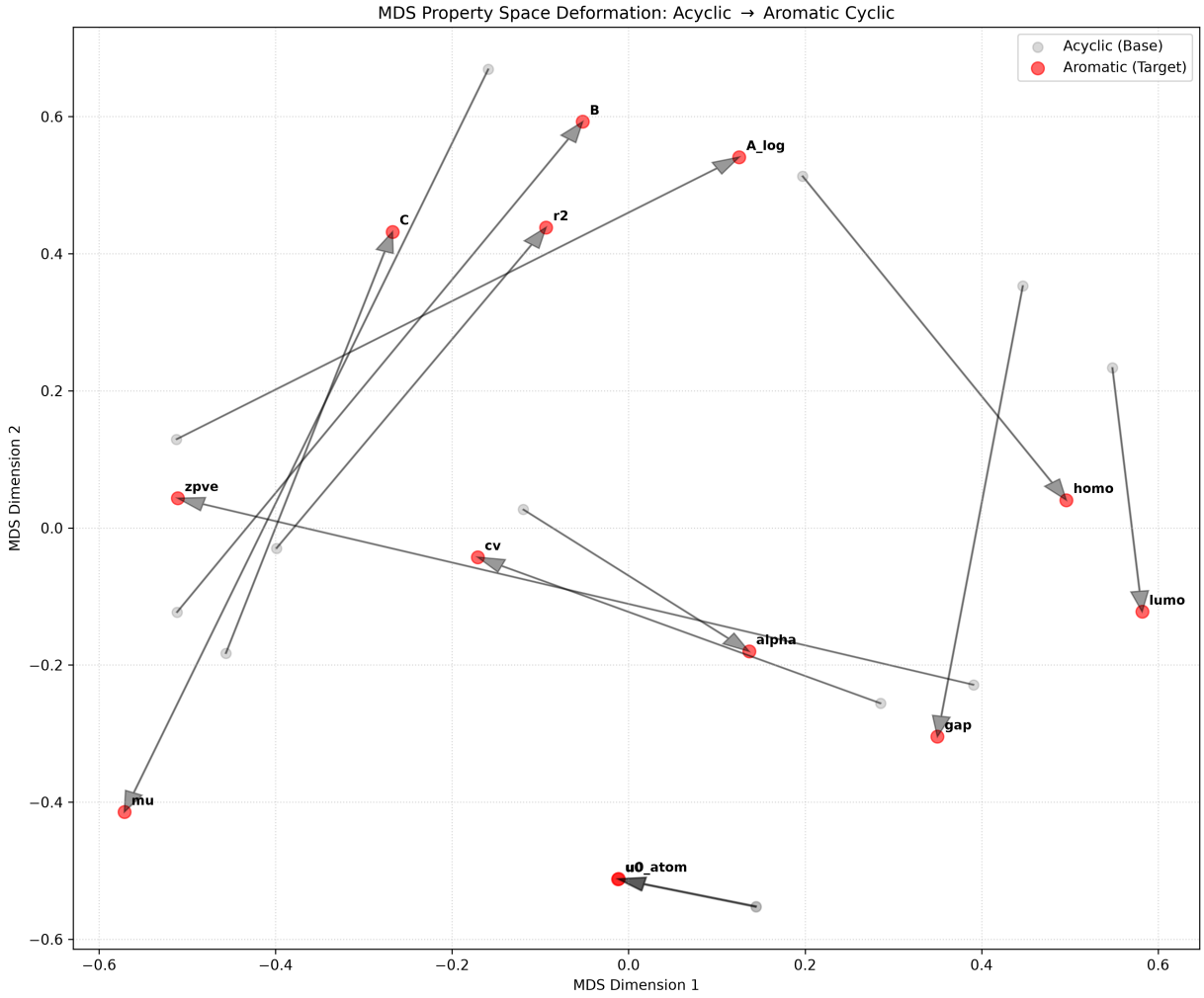

Figure 3: Multi-Dimensional Scaling (MDS) and Procrustes analysis illustrating the deformation of the molecular property space during the transition from acyclic (gray circles) to aromatic cyclic (red circles) architectures. Arrows denote property trajectories, visualizing how aromaticity reconfigures the joint structural manifold of the molecules. As depicted, the dipole moment (μ) exhibits the largest Procrustes deformation, indicating that its coupling to the broader property space is most profoundly altered by the presence of aromatic rings. Furthermore, the significant repositioning of electronic orbital energies, specifically *homo* and *gap*, reflects their shift toward higher network centrality, where they emerge as hubs mediating other physical properties in aromatic systems. This visualization demonstrates that molecular properties are not independent targets, but rather form a structured state space that is fundamentally rewired by architectural features like delocalization.

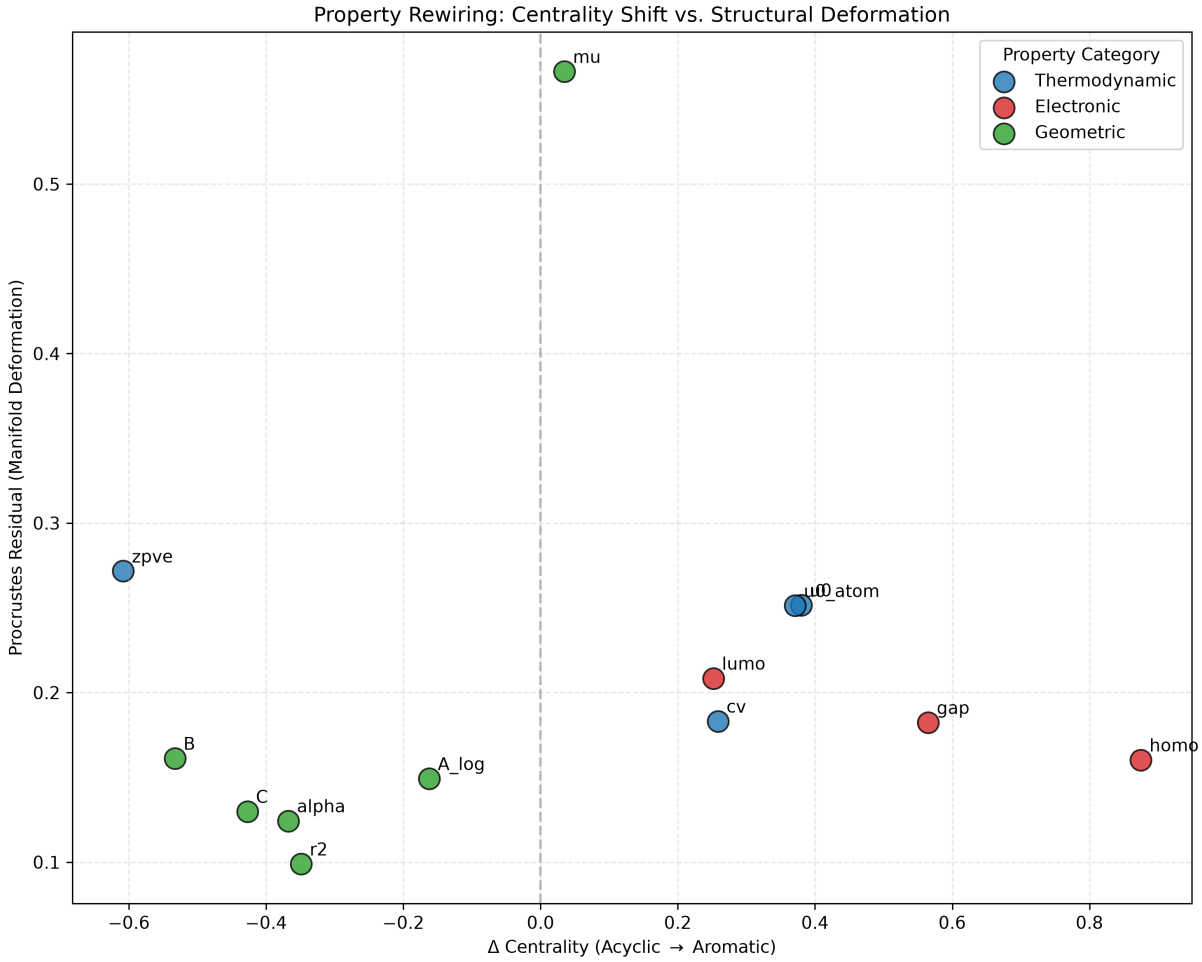

Figure 4: Scatter plot illustrating the structural rewiring of molecular properties during the transition from acyclic to aromatic architectures. The horizontal axis displays the change in network centrality, while the vertical axis shows the manifold deformation quantified by the Procrustes residual. The dipole moment (μ) exhibits the highest Procrustes residual, indicating that its coupling to the rest of the property space is profoundly altered by aromaticity. Furthermore, electronic properties, particularly *homo* and *gap*, show the largest increases in centrality. This demonstrates that in aromatic systems, electronic states become central hubs that mediate other physical properties, a role they do not hold in acyclic molecules.

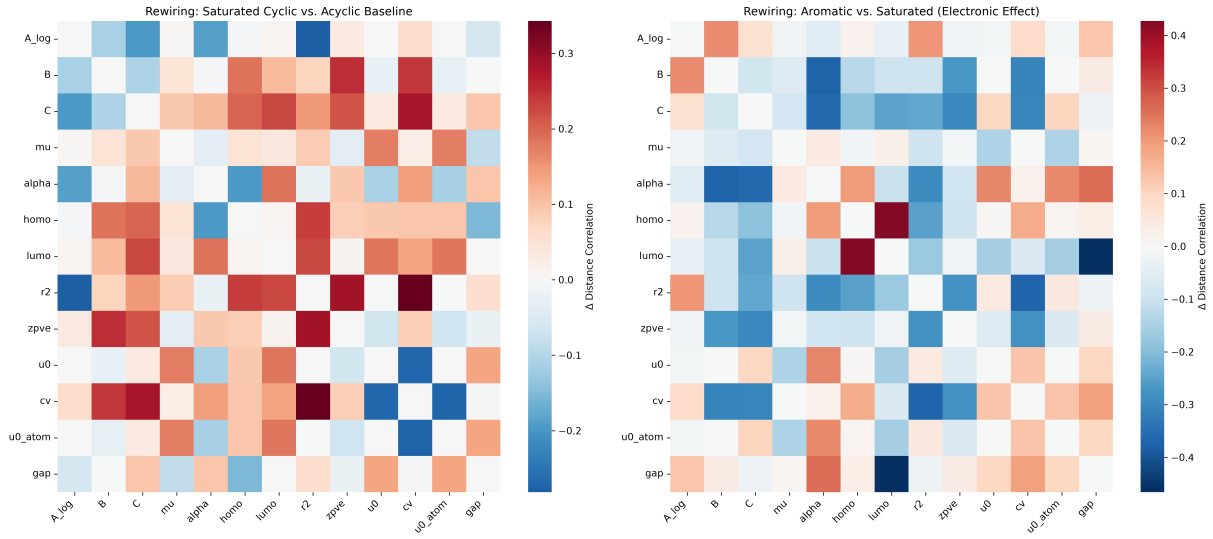

Figure 5: Difference heatmaps of distance correlation ($\Delta dCor$) illustrating the structural rewiring of molecular properties across architectural families. The left panel compares saturated cyclic to acyclic baselines, while the right panel highlights the profound deformations introduced by aromaticity relative to saturated structures. The transition to aromatic rings drives specific coupling changes, notably an electronic rewiring where the *lumo-gap* relationship drops significantly ($\Delta dCor = -0.4660$) and the *homo-lumo* coupling increases ($\Delta dCor = +0.4273$). Furthermore, aromaticity weakens the geometric sensitivity between rotational constants (B, C) and polarizability (α), and disrupts the tight vibrational link between electronic extent (r^2) and heat capacity (c_v). Ultimately, these shifts reveal that molecular architecture acts as a primary rewiring agent, fundamentally reorganizing the joint property space by decoupling classical geometric-vibrational axes and altering electronic hubs.

tions.

Ultimately, these findings establish that the molecular property space is not a static, flat collection of targets, but a structured manifold that dynamically deforms across discrete chemical boundaries. Recognizing and statistically quantifying these scaffold-specific topological shifts provides a more rigorous understanding of fundamental molecular behavior. Future computational chemistry and machine learning models could be significantly improved by explicitly accounting for this intrinsic rewiring, moving beyond independent target prediction to capture the true joint physical structure of chemical space.